UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería de Sistemas y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

DETECCIÓN DE VEHÍCULOS EN ENTORNOS MULTI-CÁMARA UTILIZANDO INFORMACIÓN CONTEXTUAL

Juan Enrique de Santiago Rojo
Tutor: Rafael Martín Nieto
Ponente: José María Martínez Sánchez

Junio 2018

DETECCIÓN DE VEHÍCULOS EN ENTORNOS MULTI-CÁMARA UTILIZANDO INFORMACIÓN CONTEXTUAL

Autor: Juan Enrique de Santiago Rojo
Tutor: Rafael Martín Nieto
Ponente: José María Martínez Sánchez

Video Processing and Understanding Lab

Departamento de Tecnología Electrónica y de las Comunicaciones

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Junio 2018

Trabajo parcialmente financiado por el gobierno español bajo el Proyecto TEC2014-53176-R (HA-Video)

Resumen

Dada la amplia demanda que actualmente existe en el área de la video-seguridad, se ha producido un aumento en el número de investigaciones que se derivan de este campo. En particular, con el fin de facilitar el control de los parkings, se presenta un sistema multi-cámara para la detección de vehículos y su correspondiente asignación a las plazas ocupadas por los mismos dentro del aparcamiento, obteniendo una visión de aquellas plazas que están disponibles para ser ocupadas. Gracias a este sistema, se puede sustituir con el uso de la visión artificial el método habitual de instalación de sensores de inducción o de peso y movimiento, los cuales encarecen considerablemente el despliegue y gestión de estos aparcamientos.

De cara a proporcionar sistemas cada vez más eficientes, son muchos los algoritmos que han surgido para la detección de objetos y sus características. En concreto en este trabajo se hace uso de dos algoritmos de detección de objetos, *Deformable Parts Model* (DPM) y un detector de regiones mediante redes neuronales convolucionales, *Faster Regions with Convolutional Neural Network* (Faster R-CNN). Ambos han sido probados en trabajos previos.

Con el fin de mejorar estos sistemas de detección de vehículos, se propone la integración de la información del entorno captada por un conjunto de cámaras, de manera que la fusión de esta información proporciona un mayor rendimiento a la hora de realizar detecciones dentro del aparcamiento. Gracias a ello, detecciones que no se pueden llevar a cabo desde un punto de vista por posibles oclusiones o grandes distancias son ahora posibles puesto que se completa esta información con la que otro punto de vista (u otros) proporciona.

Palabras clave

Parkings, detección de vehículos, información contextual, homografías, fusión multi-cámara, frame.

Abstract

Due to the high demand present nowadays in the video-surveillance area, the number of researches related to this field has been increased. In particular, with the purpose of making parking access easier, a multi-camera system is proposed for the vehicle detection and its corresponding mapping into the parking spots of a parking lot. Therefore, we will obtain the knowledge of which parking spots are free or occupied. Thanks to this system, the use of the common deployed sensors can be substituted by computer vision, which also reduces system costs.

In order to make more efficient systems, many algorithms have been developed for object detection. In this work, two object detection algorithms have been used, Deformable Parts Model (DPM) and a Faster Regions with Convolutional Neural Network (Faster R-CNN). Both tested in previous works.

The main goal in this work is to improve a vehicle detector system by fusing the information captured from the environment, from a set of cameras. The detections performed in the parking are improved, so some detections which were not possible to detect from one of the cameras now obtained thanks to another camera´s information. This helps to get detections from those objects that are far away or occluded as different points of view are taken into consideration.

Keywords

Parking, vehicle detections, contextual information, homographies, multi-camera fusion, frame.

Agradecimientos

Durante estos años que ha durado mi paso por esta universidad, han sido muchas las personas a las que les debo agradecer algo. En primer lugar, quiero agradecer a mi familia que me apoya y acompaña siempre, me han hecho, no solo lo que soy, si no lo que seré. A Claudia que desde el primer momento me ha ayudado y hecho mejorar, gracias por estar siempre. Indudablemente debo agradecer a Rafa, mi tutor, quien sin su ayuda y buen tutelaje este Trabajo Fin de Grado no habría sido posible, gracias por confiar en mí para este trabajo. No puedo olvidarme de todos mis compañeros que han formado y forman parte de este intenso recorrido académico, Álvaro, Alejandro, Rodrigo y Paula, a todos, muchas gracias por ser tan buenos compañeros y mejores amigos.

A todos, muchas gracias.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Actualmente, la detección de objetos en entornos video-monitorizados es una tarea que ha generado un gran interés en la comunidad científica. Hay muchas aproximaciones que buscan ofrecer una solución a este problema, tanto para escenarios controlados como para aplicaciones específicas de vigilancia. Los sistemas de video-monitorización intentan extraer información de manera automática a partir de la secuencia de video y, además, procuran generar una descripción de la escena lo suficientemente útil como para que la interacción con el humano resulte eficiente.

Existen, dentro del campo de la visión artificial, una gran variedad de algoritmos de segmentación, detección de objetos, reconocimiento de eventos, etc., que son usados en estos sistemas de video vigilancia. Gracias al entrenamiento de dichos sistemas, de los que obtenemos gran cantidad de datos, se consigue modelar cualquier tipo de objeto y su detección de forma automática.

Los parkings de vehículos son un servicio ampliamente utilizado en el que cada año se realiza una gran inversión. La gestión de estos aparcamientos es muy costosa y en muchos casos compleja, especialmente en el caso de aquellos lugares formados por muchas plazas, como aeropuertos o grandes zonas comerciales. Resolver este problema utilizando la visión artificial, en combinación con los sistemas de video vigilancia presentes en estos aparcamientos para monitorizar las plazas que están ocupadas, promete una serie de ventajas sobre los sensores intrusivos.

La motivación fundamental de este Trabajo Fin de Grado, TFG, es la de contribuir al desarrollo y perfeccionamiento del proceso de combinación de la información generada por cada una de las cámaras. Para que, con esto, se puedan mejorar las detecciones llevadas a cabo por nuestro sistema. Se debe aclarar cuál es la fuente fundamental de la información que se va a fusionar en cada cámara, esta es la información contextual. Se conoce como información contextual, a aquella información disponible inicialmente del escenario (por ejemplo, distancias entre elementos del entorno de grabación, posición de las cámaras con respecto a un sistema de referencia, etc).

1.2 Objetivos

Los objetivos que podemos considerar para este TFG comienzan por analizar y comprender el estado del arte respecto de los sistemas de gestión automática de vehículos, además de los sistemas de detección de objetos como base fundamental de la detección de vehículos. Por supuesto, dado que este TFG realiza una mejora en una parte del sistema concreto del que se parte, la comprensión y análisis de dicho sistema resulta fundamental y obligada como objetivo. Se debe llevar a cabo esta mejora desarrollada en la motivación, una evaluación rigurosa de la mejora con respecto al punto de partida y verificar que dicha mejora cumple los objetivos deseados.

1.3 Organización de la memoria

La memoria consta de las siguientes secciones:

1. Motivación, objetivos y organización de la memoria
2. Estado del arte
3. Diseño y desarrollo
4. Experimentos y resultados.
5. Conclusiones y trabajo futuro
6. Referencias

2 Estado del arte

Este capítulo tiene como objeto proporcionar una visión global del trabajo previo en el conjunto de esta área y las áreas derivadas. El capítulo se divide en cuatro secciones, en la primera de las secciones se hace una clasificación de los detectores de objetos, en la segunda sección se clasifican las bases que conforman los detectores de vehículos, en la tercera sección se desarrollan los algoritmos de detección usados en el sistema base del que partimos y en la cuarta sección se estudia el trabajo relacionado con el objetivo principal de este TFG.

2.1 Detección de objetos

Dentro del campo de la visión artificial, la capacidad para detectar y diferenciar objetos es, sin duda, uno de los mayores retos. Existen varios sistemas de detección de objetos, sin embargo, podemos afirmar que, en general, poseen una estructura similar y compartida.

Según [1] se plantean dos aproximaciones de la detección de objetos, aquella basada en una cierta segmentación de la escena en un primer plano (objetos) y en el fondo; y aquella basada en una búsqueda exhaustiva. Existen, además, otras aproximaciones que tratan de combinar las dos anteriores. Para cualquiera de estas, el resultado termina por ser la localización y dimensionado del objeto.

A continuación, se detalla cada una de estas dos aproximaciones:

- Segmentación: mediante este método se trata de detectar objetos en movimiento, mediante la comparación del *frame* actual con un *frame* de referencia, y limitar los resultados para obtener los objetos de interés.

- Búsqueda exhaustiva: esta otra técnica que nos permite obtener la posición inicial del objeto consiste, normalmente, en el escaneado completo de la imagen buscando similitudes con el modelo de objeto que se busca. Gracias a esta técnica se obtiene

un mapa de confianza denso. Para obtener la detección individual, esta aproximación debe buscar los máximos locales.

- Segmentación y búsqueda exhaustiva: como última aproximación se plantea una combinación de las dos anteriores de manera que se refuerzan sus puntos fuertes y se aplaquen sus debilidades.

2.2 Detección de vehículos y plazas ocupadas o vacías.

A continuación, se detallan los distintos enfoques de sistemas de detección que, tras haber estudiado el estado del arte [2], se han determinado como principales en la detección de objetos:

- Sistemas basados en segmentación: este método es empleado para dividir la imagen en regiones separadas que, en teoría, se corresponden con los distintos objetos del mundo real. Más concretamente, lo que este proceso propone es la etiquetación de todos los píxeles que conforman la imagen, gracias a esto, aquellos píxeles cuya etiqueta es la misma se pueden agrupar puesto que comparten características o propiedades, tales como el color y la textura. Junto con esto y aplicando el procedimiento de etiquetado, pero para las regiones contiguas que poseen características muy distintas, se consigue la discriminación y la localización de los objetos respecto del fondo de la imagen.

- Sistemas basados en clasificación por parches en las plazas, estos sistemas usan técnicas de clasificación de *machine learning* entrenadas con entornos etiquetados previamente. Algunos de los trabajos más representativos son [3]-[12]. El algoritmo propuesto en [4] usa una combinación de puntos característicos de un coche detectados y una clasificación por histograma de color para detectar las plazas libres. Sin embargo, presenta problemas con las oclusiones presentes.

- Sistemas basados en detectores, en nuestro caso detectan vehículos. Estos usan un algoritmo de detección con el fin de detectar los vehículos y poder mapearlos en las plazas del parking. Dentro de esta categoría, el trabajo más representativo es [13].

En este trabajo se describe un detector de coches basado en redes neuronales convolucionales. En él, tras el entrenamiento de la red, se propone buscar en la imagen completa de un parking y así identificar si existe un coche, mediante el uso de una ventana deslizante. Gracias a los avances y evolución de estos detectores de objetos, han sido viables estos sistemas basados en detectores. Especialmente relevante ha sido la aportación realizada por [14], [15] y [16].

2.3 Algoritmos de detección

A continuación, se detallan tres algoritmos de detección a tener en cuenta para el desarrollo de este trabajo.

2.3.1 *Aggregate Channel Features* (ACF)

Este algoritmo de detección, *Aggregate Channel Features*, ACF, se detalla en [17] realiza búsqueda exhaustiva, utilizando un modelo de objeto holístico. Fundamentalmente, en un principio se calculan diferentes canales, $C = \Omega(I)$, para una imagen de entrada I, para llevar a cabo un suavizado por sub-muestreado de la suma de todos los bloques de píxeles que encontramos en C. Por último, mediante el uso del impulso, se entrena y combinan los árboles de decisión sobre las características y así poder distinguir el objeto del fondo, esto se hace usando ventanas deslizantes multi-escala.

2.3.2 *Deformable Parts Model* (DPM)

En [18] se describe un sistema de detección de objetos que representa objetos altamente variantes usando mezclas multi-escala de modelos de partes deformables. Estos modelos son entrenados usando un proceso discriminante que tan solo requiere cajas (*bounding boxes,* BB) que rodeen los objetos del conjunto de las imágenes. Aunque ofrecen un buen ámbito de trabajo para la detección de objetos, ha sido difícil establecer su valor en la práctica. Para conjuntos de datos difíciles, estos modelos DPM no rinden lo suficiente frente a otro más sencillos como *RIGID TEMPLATES* o *BAG-OF-FEATURES*.

Este modelo está formado por un filtro de raíz gruesa, que llega a cubrir todo el objeto, considerándolo como componente principal, y otros filtros *P* más pequeños que cubren aquellas partes más pequeñas del objeto. Este modelo de detección es un algoritmo basado en la búsqueda exhaustiva, aunque con un modelo basado en partes.

2.3.3 FASTER R-CNN

Teniendo en cuenta el estado del arte [16], los detectores basados en redes convolucionales hacen uso del aprendizaje profundo para la extracción y la selección de aquellas características con las que es capaz de discriminar los objetos a detectar.

Este algoritmo es una variación de una versión anterior llamada *Fast Region-based Convolutional Network,* Fast R-CNN y del algoritmo original *Region-based Convolutional Network,* R-CNN [15].

Simplemente mencionar algunos detalles acerca de R-CNN, este algoritmo se encuentra dividido en cuatro etapas. En la primera de ellas se toma una imagen como elemento de entrada al sistema. En la segunda etapa del detector se generan las denominadas regiones de interés, sobre las cuales se aplicará el detector. Es en la tercera etapa donde se extrae el vector de características con longitud fija para cada una de las regiones mencionadas anteriormente. Y en último lugar, la cuarta etapa consiste en un conjunto de clases lineales SVM. Se puede afirmar que el coste computacional de este algoritmo es bastante alto, haciendo de él un algoritmo lento.

Gracias a la utilización del algoritmo FAST R-CNN, se consiguen resolver los problemas que el uso de redes convolucionales genera [15]. La manera de hacer esto es mediante la introducción de una imagen y unas ciertas regiones de interés en una red convolucional para su clasificación por capas de agrupación máximas y de esta forma crear el mapa de características. Se extraen los vectores de características para cada región de interés obtenida en el paso descrito anteriormente. El tamaño de este vector es fijo. Estos vectores se añaden uno a continuación de otro y se crea una secuencia que podría decirse está totalmente conectada, *FC*; dicha secuencia es dividida en dos capas iguales, las cuales producen: una de las capas, estimaciones de probabilidades *softmax* sobre *k* clases de objetos más una clase de tipo fondo y la otra capa, genera 4 valores reales, los cuales codifican las posiciones de la caja delimitadora por cada *k* clases.

En definitiva, el algoritmo de detección propuesto en [16], Faster R-CNN, está compuesto por dos módulos. El primero es una red convolucional pura profunda que procesa las regiones y el segundo es un detector Fast R-CNN. Todo el sistema conforma una red única de detección de objetos.

Se denomina al primero de estos módulos como *Region Proposal Network*, RPN, el cual toma una imagen de cualquier tamaño como entrada y devuelve a la salida un conjunto de propuestas de objetos definidos por rectángulos, cada uno de ellos con una puntuación objetiva. De forma que se generen estas regiones, aplica al mapa de características una serie de pequeñas redes a través de ventanas deslizantes. Se le asigna a cada una de estas ventanas deslizantes una característica de menor dimensión. Con esto, alimenta dos capas hermanadas y completamente conectadas, una para la clasificación de las cajas *cls* y la otra para la regresión de las cajas *reg*. De forma simultánea, en cada una de las localizaciones de la ventana deslizante, se calculan las posibles regiones de interés. Como máximo se establecen *k* regiones. En la capa *reg* se encuentran las coordenadas de las posibles regiones y en la capa *cls* se encuentran las probabilidades de que dichas regiones sean o no un objeto. Con todo ello, habiendo determinado, según lo expuesto, las regiones de interés, el algoritmo Fast R-CNN es ahora aplicado dando paso a la detección. Se puede afirmar, en definitiva, que es RPN quien le dice a Fast R-CNN dónde debe buscar.

2.4 Mejora de las detecciones utilizando información contextual

Centrándonos en el objetivo de este TFG, analizaremos el trabajo relacionado con la mejora de las detecciones a partir de la utilización de la información contextual. Sabiendo que la información contextual es toda aquella que rodea al escenario grabado (distancia entre objetos detectados y cámaras, posición de las cámaras, etc.) se propone en [19] hacer uso de ésta, de tal forma que se pretende conseguir una mejora en las detecciones, transfiriendo información de unas cámaras a otras y combinándola.

Puesto que la información que proporciona una sola cámara puede resultas escasas, para poder monitorear un área amplia, resulta necesario el uso de información complementaria, proporcionada por otras cámaras. Ese hecho se acentúa cuando existen en el escenario oclusiones o condiciones climatológicas desfavorables para la detección de objetos. Las tecnologías principales para el uso de multi-cámaras se explican en [19].

Tal y como se describe en [19], puesto que este trabajo utiliza una modificación de éste método descrito, las detecciones de una cámara se transfieren a la otra, en vez de proyectar todas las detecciones en el plano común, tal y como proponen [21] y [22]. En tal caso, el plano común se usará para obtener la información de cada punto de vista de las cámaras y para transferir y combinar las detecciones de objetos de una cámara a otra. Se plantea este método en un *dataset* de personas, en los que las posiciones de las personas se aproximan mediante cilindros, manteniendo así el volumen ocupado por éstas.

2.4.1 Técnica de transferencia de información contextual para personas

El objetivo de la técnica desarrollada es el de usar los bounding boxes (BB) de las detecciones de una cámara y transferirlos al punto de vista de la otra cámara. Puesto que las proyecciones en el plano común no se corresponden espacialmente con la posición del objeto detectado, la transferencia entre cámaras debe ser corregida. A continuación, se detalla la técnica desarrollada en [19], siendo una de varias, en la cual se fundamenta este trabajo:

1. Primero, se proyecta sobre el plano común el segmento de la base del BB correspondiente a la detección de la persona (ver Figura 2-1). Tal y como se describe más adelante, el plano común es obtenido mediante homografía.

Figura 2-1. Proyección del segmento base.

2. Usando la proyección del segmento, se define una circunferencia de forma que el segmento forme uno de los lados de un cuadrado inscrito en esta circunferencia (ver Figura 2-2).

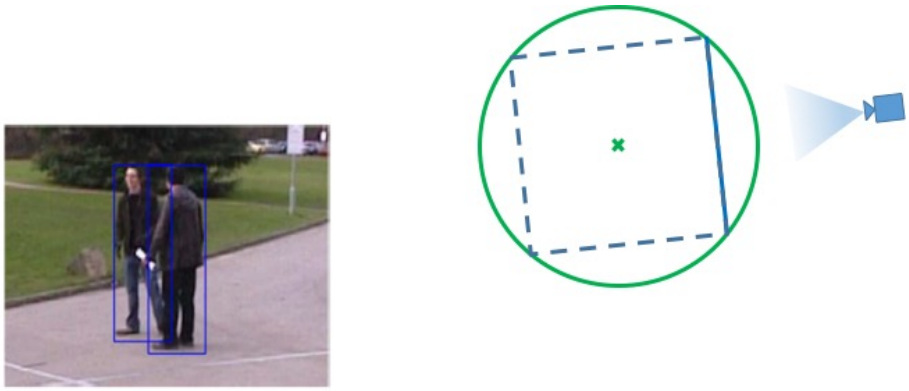


Figura 2-2. Definición de la circunferencia y cuadrado inscrito.

3. Para definir la base del BB que será transferido a la otra cámara, el cuadrado inscrito se rota, según un ángulo tal que, el lado más cercano sea perpendicular a la línea que une la cámara con el centro de la circunferencia definida anteriormente (ver Figura 2-3). Este nuevo segmento se corresponde a la proyección del segmento de la base del BB que debe ser transferido.

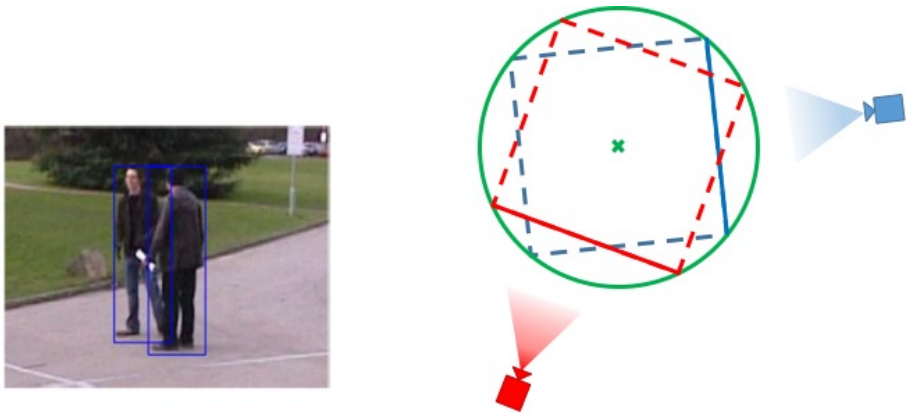


Figura 2-3. Rotación del cuadrado.

4. La altura del cilindro se estima utilizando reglas de proporcionalidad, teniendo en cuenta la altura inicial del objeto y la distancia del objeto a las cámaras. En la Figura 2-4 se ilustra el resultado obtenido; en verde, el cilindro estimado, en azul el BB original y en rojo el BB resultante.

5. Finalmente, este BB es transferido al punto de vista de la otra cámara, usando la homografía inversa.

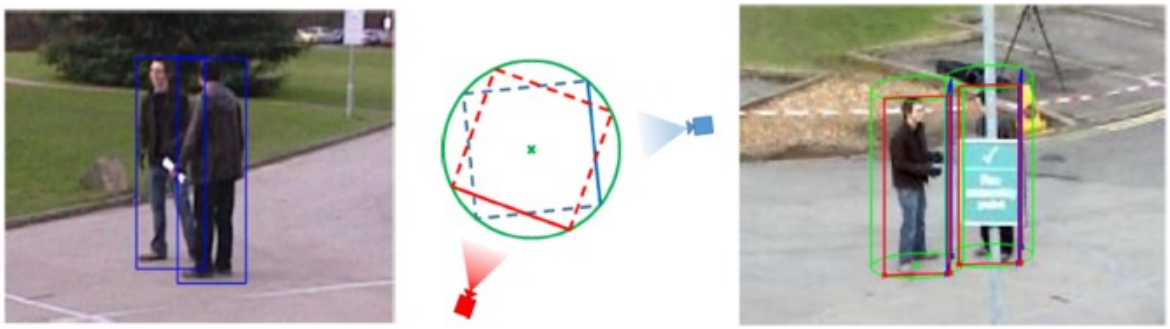


Figura 2-4. Estimación de altura y generación del *bounding box*.

Para la combinación de las detecciones, puesto que se pueden producir múltiples coincidencias de BB para la misma persona, éstas se simplifican a una sola. Las medidas usadas para determinar esta asociación son las mismas que las usadas para decidir si dos BB se corresponden con el mismo objeto. Dos BB de una detección corresponden a una persona si $rd \leq 0.5$ (*relative distance)*, que es la distancia relativa entre dos BB teniendo en cuenta una desviación del tamaño del objeto de un 25 % y el *cover* y *overlap* están por encima del 50%. El BB que mejor confianza presenta es el que se mantiene. Estas métricas se representan en la Figura 4-4 y se da más detalle acerca de éstas en la sección 4.1.3 de esta memoria, vienen definidas por [22].

Figura 2-5. Resultado de la combinación de información.

En la Figura 2-5 se ilustra un ejemplo de la combinación de las detecciones. Arriba a la izquierda, las detecciones propias de una cámara; abajo a la izquierda, las detecciones transferidas; y a la derecha, se representan en verde las detecciones propias, en azul las transferidas y el *ground truth*, GT, en rojo.

Este trabajo considera los detectores DPM y *Faster* R-CNN para determinar si el uso de esta información contextual proporciona mejores detecciones en entornos multi-cámara. Según se concluye en [19], los resultados obtenidos confirman la suposición de que la combinación de información proporciona mejores resultados que los arrojados del procesamiento de las cámaras por separado. Funcionando la técnica descrita en escenarios diferentes, para personas con distinto aspecto (de pie, sentadas, etc.) y para cualquier orientación entre las distintas cámaras, gracias a la asunción volumétrica de las personas descrita más arriba.

3 Desarrollo

3.1 Introducción

El sistema multi-cámara propuesto en [2], está basado en un procesamiento paralelo de cada cámara seguido de la combinación de sus resultados individuales. El conjunto de cámaras toma imágenes que se procesan por un detector de objetos usando un modelo de coches previamente entrenado. Con esto, cada detección es procesada y mapeada en un plano común del parking para monitorizarlas. Para tener en cuenta el volumen de los objetos detectados es necesario realizar una corrección de perspectiva. Luego, usando una conversión mediante homografía, es posible mapear las detecciones en las posiciones determinadas por la matriz de ocupación. Por ultimo, la información que se obtiene del procesamiento de cada una de las cámaras es combinada de forma que se genera una matriz final con los espacios ocupados y vacíos del parking.

En el punto de la corrección de perspectiva se hará una transferencia de la información de una cámara a otra para, de esta forma, cumplir el objetivo de mejorar las detecciones. Esto se lleva a cabo en el punto en el que se procesan y mapean las detecciones en el plano común. Utilizaremos la técnica desarrollada en [19], descrita en la sección 2.4.1, aunque con algunas variaciones. De esto se obtienen nuevos BB que son transferidos al punto de vista de la otra cámara y combinados con las detecciones de esa cámara, complementándose.

3.2 Sistema de detección

A continuación, se detallan los bloques mediante los cuales nuestro sistema de detección de vehículos hace efectivo el procesamiento de los *frames* tomados por cada cámara y devuelve la especificación de la ocupación del parking. El diagrama de bloques del sistema de detección se muestra en la Figura 3-1.

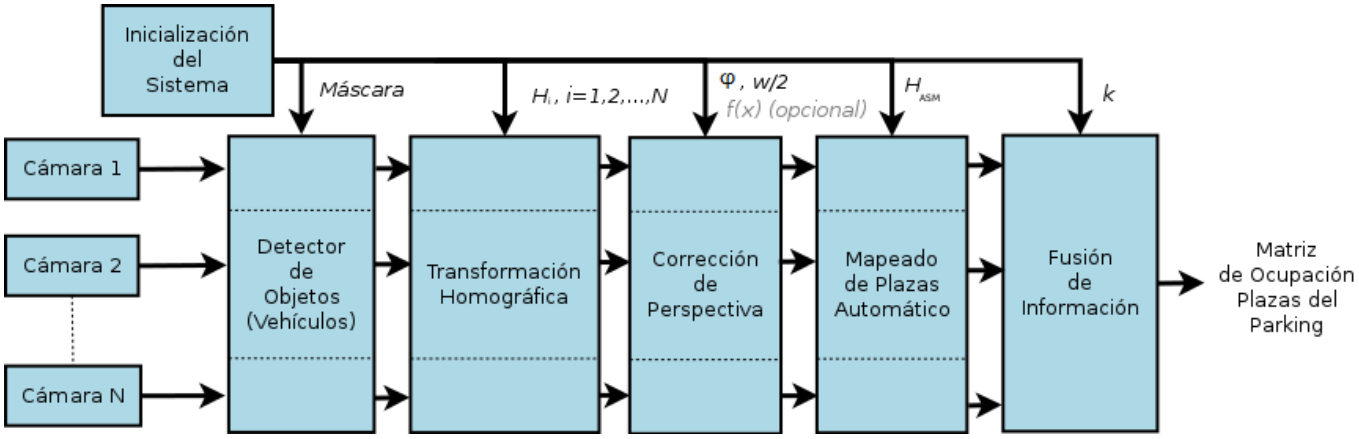


Figura 3-1. Diagrama de Bloques del Modelo original.

Cada uno de los bloques se explicarán a lo largo de esta sección. Las cámaras generan los *frames* a procesar (izquierda), el detector de objetos (vehículos) genera las detecciones y devuelven los BB, mediante dos algoritmos: DPM y Faster R-CNN; se aplica una transformación mediante homografía y una corrección de perspectiva a los BB, la información contextual proporcionada por el conjunto de cámaras es fusionada y como resultado se obtiene una matriz de ocupación del parking. Este es el sistema de detección propuesto en [2].

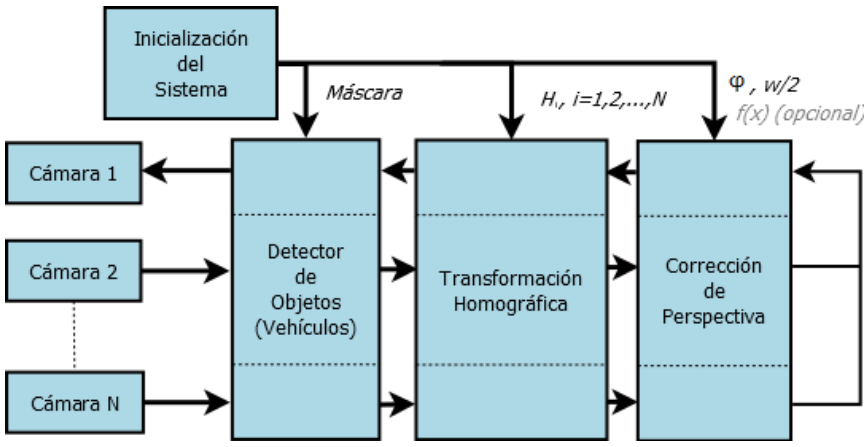


Figura 3-2. Diagrama de Bloques del Modelo nuevo.

Este nuevo diagrama de bloques (ver Figura 3-2) es el que se obtiene de este trabajo y se explicarán en la sección 3.3 los cambios realizados respecto del original (ver Figura 3-1). Tanto el mapeado de plazas automático como la fusión de información desaparecen ya que no son bloques que se hayan utilizado en el desarrollo de este trabajo. En el bloque del detector de objetos del diagrama del nuevo modelo, es donde concluye la combinación de la información contextual, comenzando después del bloque que corrige la perspectiva.

3.2.1 Detector de vehículos

Este detector de objetos (vehículos) propuesto se inicializa con el modelo de vehículo y, además, cada *frame* a analizar es enmascarado mediante una ROI (*Region Of Interest*). De esta forma eliminamos posibles detecciones de áreas ajenas a la que será monitorizada por nuestras cámaras. Se muestra un ejemplo de esto en la Figura 3-3.

a **b** **c**

Figura 3-3. Ejemplo de máscara utilizada.

Ejemplo de (a) la máscara, (b) *frame* de entrada de la cámara 1 y (c) *frame* enmascarado.

El detector de vehículos es el que recibe cada uno de los *frames* que se van a procesar de cada cámara y, usando uno de los algoritmos de detección descritos anteriormente, devuelve o genera un BB, rectangular, correspondiente a cada uno de los vehículos detectados. Estos algoritmos de detección propuestos son: Faster R-CNN, una variación más eficiente de la versión anterior R-CNN y Fast R-CNN en cuanto a coste computacional y rendimiento. Y el segundo de estos detectores es *Deformable Parts Model* (DPM). El cual se basa en una búsqueda exhaustiva y un modelo basado en partes. Este se ha utilizado para ver el comportamiento del sistema cuando no se usa un detector basado en *deep-learning*. Con esto se demuestra la robustez de estos detectores.

3.2.2 Transformación homográfica

Para poder llevar a cabo la combinación de la información debemos trabajar con ésta en un plano común (ver Figura 3-4) que se obtiene gracias a las homografías y sus propiedades.

La matriz de homografía usada, H_i, se obtiene usando cuatro puntos de cada punto de vista de las cámaras. Cabe decir que, estos puntos no tienen porqué ser los mismos en cada punto de vista, sin embargo, deben estar asociados con uno dentro del plano común. Dado su alto coste computacional, la homografía solo se aplica a cada punto medio de la base del BB y no a toda la imagen. Tras ello obtenemos un punto a la salida por cada uno de los vehículos detectados. Más adelante, en la sección 3.3.1 de este trabajo, se detallará el procedimiento mediante el cual fusionamos la información contextual proporcionada por la cámara contraria a la que en ese momento se esté procesando.

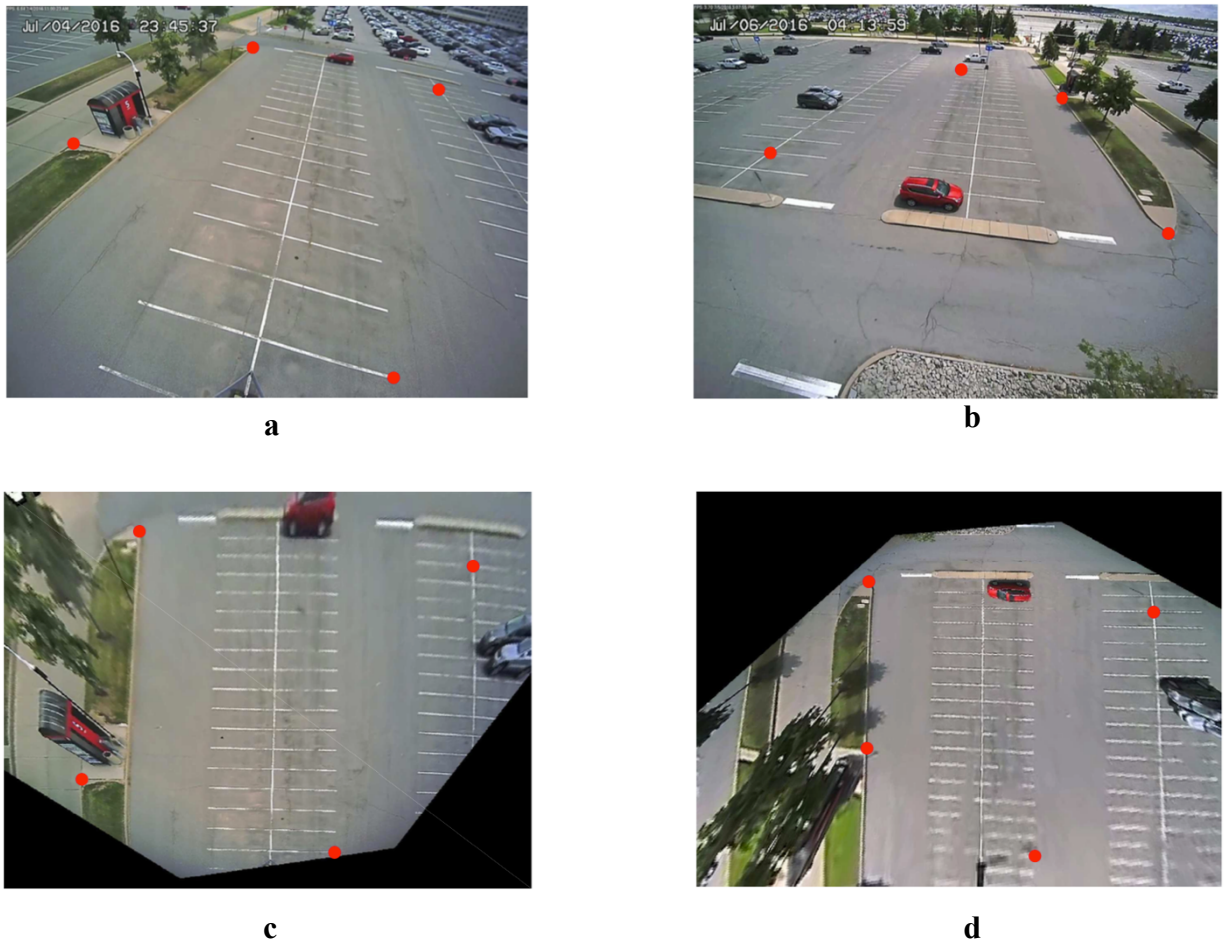


Figura 3-4. Puntos de vista de la transformación por homografía.

(a) y (b) muestran los puntos de vista originales y (c) y (d) la vista cenital (plano común).

3.2.1 Corrección de perspectiva y volumen

El detector plantea dos correcciones, la primera corrección actúa sobre la posición del vehículo que puede sufrir una distorsión debido al volumen de éste y la segunda corrección, actúa sobre una posible distorsión introducida por la lente del objetivo de la cámara. Para poder realizar la primera de las correcciones, se utiliza el ángulo que se forma entre las líneas de la plaza y el ángulo del punto de vista de la cámara. De esta manera, se sitúa el vehículo en el centro de la plaza en la que este se encuentra aparcada.

En la Figura 3-5 se muestra un diagrama de la corrección, *A* es el punto medio de la base de la proyección en el plano común, *B* es el punto final después de la corrección, φ es el ángulo entre el punto de vista de la cámara y las líneas del parking y $\frac{w}{2}$ es la mitad de la longitud media de un vehículo.

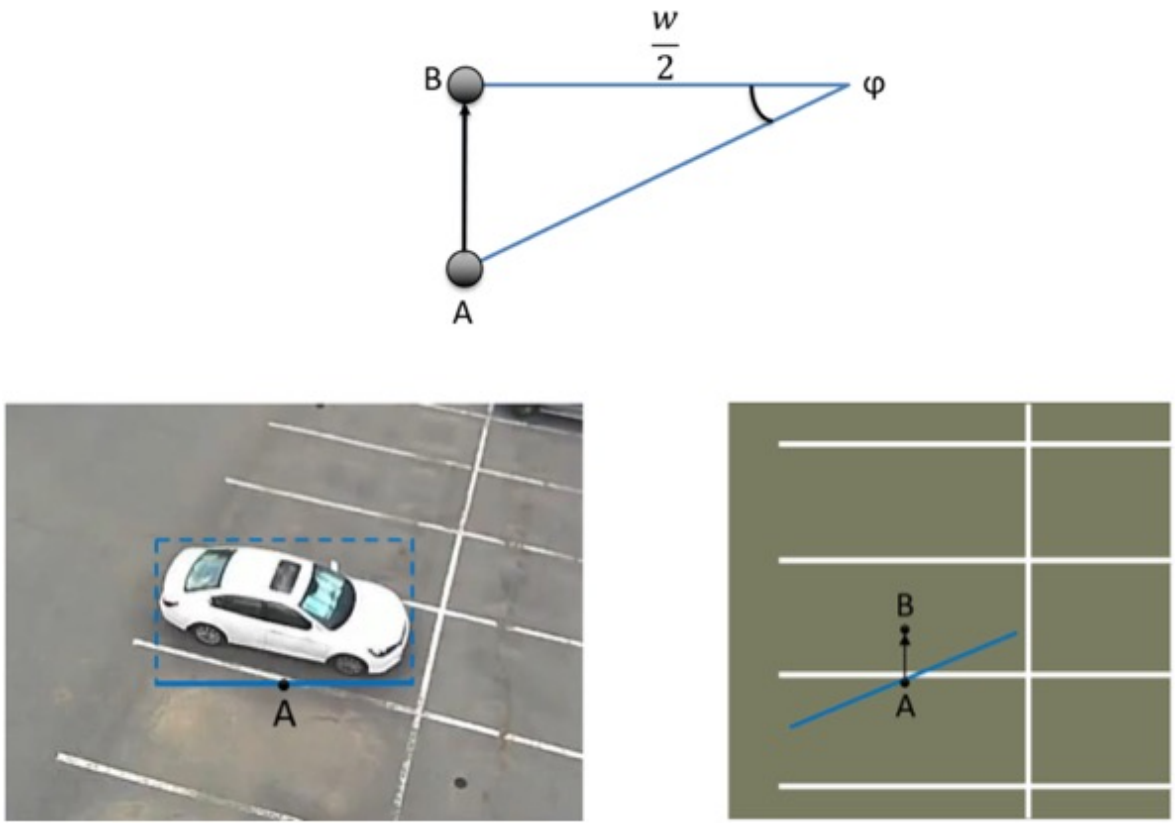


Figura 3-5. Corrección de perspectiva.

Respecto de la corrección de la lente, esta distorsión puede afectar a la precisión de la homografía, provocando errores e imprecisiones a la hora de mapear los puntos. De no existir los parámetros que determinan la distorsión de la lente, mediante los cuales se realiza el ajuste apropiado, se pueden ajustar los puntos usando una función lineal simple (y aproximada).

Estas correcciones se han tenido en cuenta a la hora de realizar la transferencia de la información contextual de una cámara a otra, ya que nos basamos en los puntos obtenidos de la homografía para llevar a cabo el fin de este trabajo.

3.2.2 Mapeado automático de plazas

Este bloque se basa en el uso de las propiedades de la homografía. Sin embargo, este bloque no se ha usado ni modificado durante el desarrollo de este trabajo (ver Figura 3-2). Más información se presenta en [2].

3.3 Implementación para la transferencia de información de una cámara a otra

Tal y como se mencionaba en la introducción de esta sección, este trabajo se basa en una combinación del trabajo realizado por [2] y [19].

A continuación, se explica el desarrollo realizado, conforme a lo expuesto como objetivos de este Trabajo de Fin de Grado, es decir, incluir información de contexto de la escena en un detector de vehículos y combinar las detecciones de distintas cámaras con el fin de mejorar los resultados de detección.

3.3.1 Transferencia información contextual

De manera análoga a como se plantea en [19], el objetivo de la técnica usada es el de utilizar los BB de las detecciones de una cámara y transferirlos al punto de vista de la otra cámara. Se ha tenido en cuenta que, dado que las proyecciones (de los BB detectados) en el plano común no se correspondían espacialmente con la posición del vehículo detectado, la transferencia entre cámaras debía ser corregida. Este sería el proceso detallado:

1. Partimos de lo que se obtiene de la corrección de perspectiva planteada en el sistema de detección y descrita en la sección 3.2.1. La base del BB de la detección es proyectada sobre el plano común, usando para ello su punto medio.

2. Este punto medio es proyectado sobre el centro de la plaza ocupada por el vehículo (ver Figura 3-6).

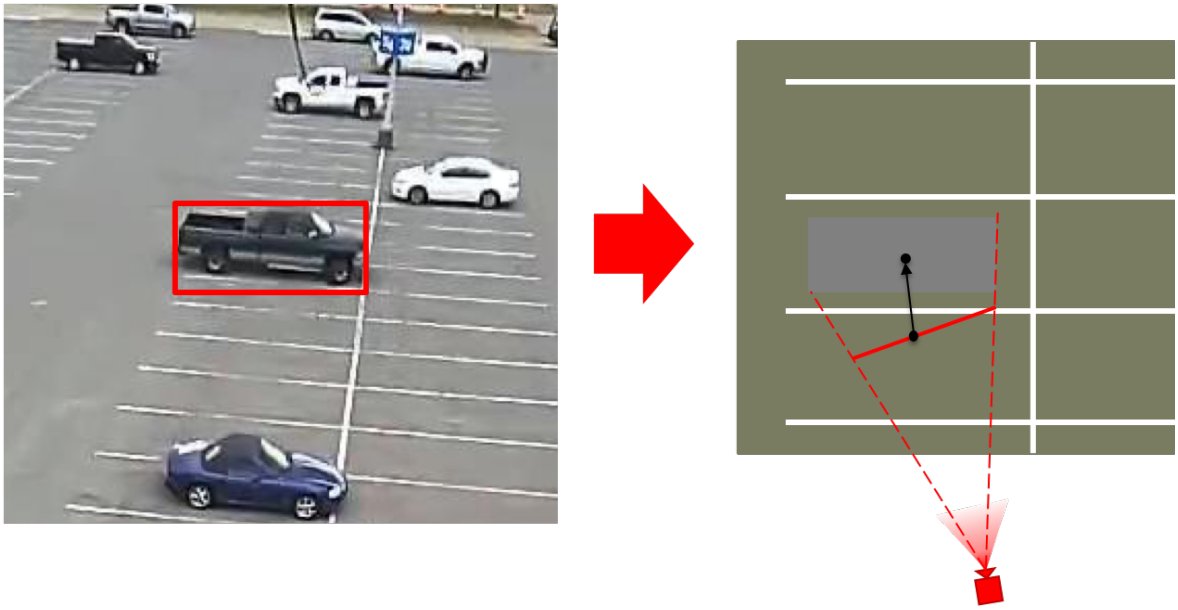


Figura 3-6. Proyección en plano común de detección.

3. Desde el centro de la plaza ocupada el proceso de transformación es aplicado a la inversa con los parámetros de la otra cámara.

4. La altura del BB transferido es calculada mediante una relación de alturas tomada como referencia. Esto se explicará con más detalle en la sección 3.3.2.

5. La dirección del ancho del BB transferido se obtiene de aplicar la transformación mediante homografía de dos puntos de correspondientes a ambos puntos de vista.

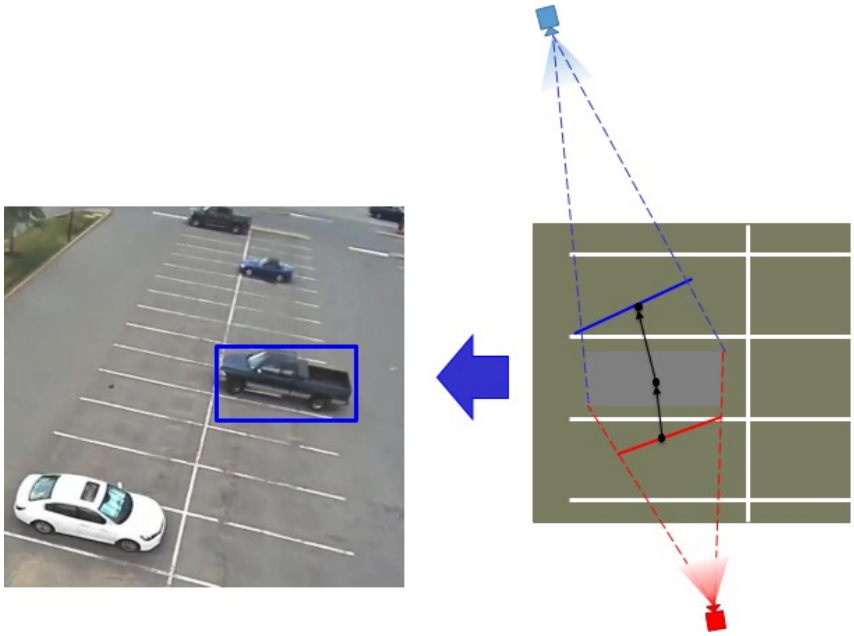


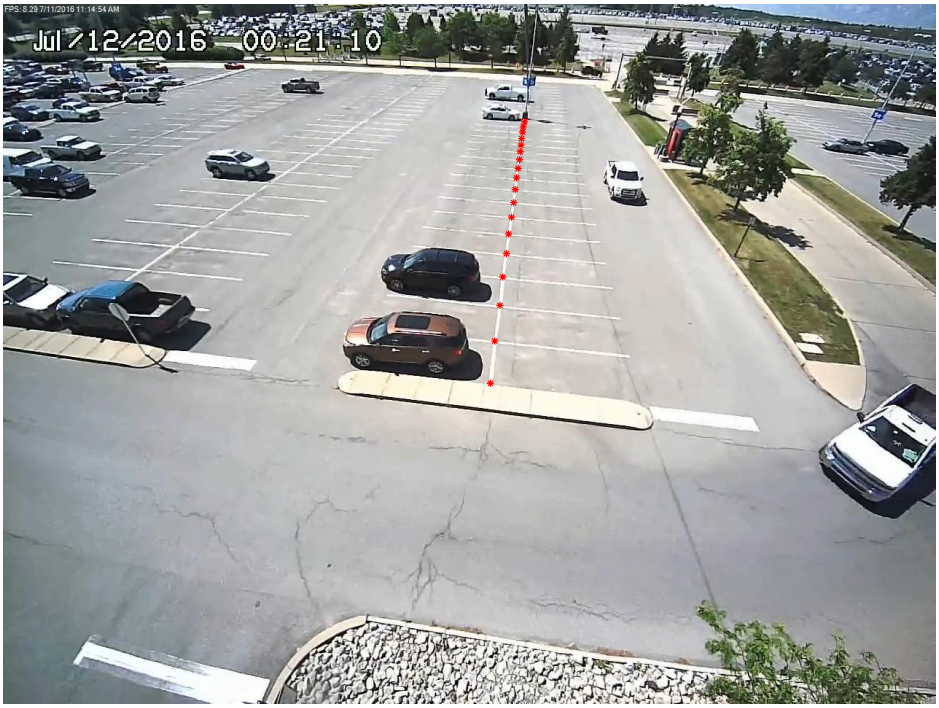
Figura 3-7. Generación del nuevo BB.

6. Por último, el BB generado es transferido al punto de vista de la otra cámara, de nuevo, mediante el uso de la homografía (matiz inversa). Esto se ejemplifica en la Figura 3-7, en azul, el BB transferido.

3.3.2 Relación de alturas. Polinomio de ajuste

Para la estimación de la altura del BB transferido se ha generado un vector de coeficientes tal que, se obtiene una relación entre las alturas de los vehículos situados en las plazas vistas desde una de las cámaras con su correspondiente, vista desde la otra cámara. De forma manual, se han marcado los puntos de la línea central del parking desde los dos puntos de vista. De la combinación de las distancias se han generado los coeficientes que, multiplicados por las alturas de los BB originales, generan las alturas de los nuevos BB.

En la Figura 3-8 se muestran las anotaciones realizadas para desarrollar lo expuesto anteriormente.



a

b

Figura 3-8. Anotaciones para obtener coeficientes de alturas.

En (a) se muestra la vista desde la cámara 1 y en (b) la vista desde la cámara 2.

En la Figura 3-9 se muestra la relación de escala para cada altura según la plaza correspondiente su aproximación a una función de tipo exponencial negativa. La función y sus parámetros son:

$$f(x) = 8{,}792 \times e^{-0{,}3182x} \quad (1)$$

Para estimar los valores de los parámetros se utiliza el método de mínimos cuadrados, donde dada la ecuación y = F(p, x) siendo *p* un vector de parámetros desconocidos, *x* una matriz o un vector de datos, *y* un vector de datos, se calcula una estimación de los valores de los parámetros del vector *p* que mejor se ajusten a la ecuación (2):

$$\min\|F(x, xdata) - ydata\|_2^2 = \min \sum_i (F(x, xdata_i) - ydata_i)^2 \quad (2)$$

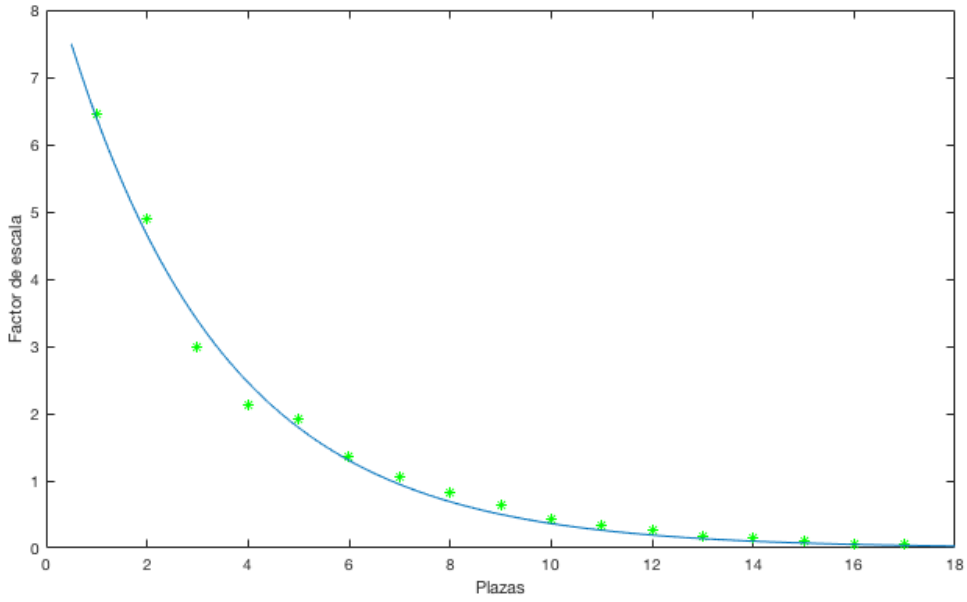


Figura 3-9. Relación de alturas y aproximación a función exponencial negativa.

3.3.3 Combinación de información de distintas cámaras

Para la combinación de las detecciones, puesto que se pueden producir múltiples coincidencias de BB para el mismo vehículo, éstas se simplifican a una sola. Las medidas usadas para determinar esta asociación son las mismas que las usadas para decidir si dos BB se corresponden con el mismo objeto. Dos BB de una detección corresponden a un vehículo si $rd \leq 0.5$ (*relative distance*), es la distancia relativa entre dos BB teniendo en cuenta una desviación del tamaño del objeto de un 25 % y el *cover* y *overlap* están por encima del 50 %. Esto se muestra en la Figura 4-4. El BB que mejor confianza presenta es el que se mantiene. Se han tomado las mismas métricas que las expuestas en la sección 2.4.1. En nuestro caso evaluamos siempre con un 50 % y fusionamos los BB tanto para un 50 % como para un 100 % del *cover* y *overlap.* Aplicando un 50 %, las detecciones lejanas de una cámara se combinan con las cercanas de la otra y viceversa, existiendo una compensación entre ambas. Si se aplica un criterio del 100 %, las detecciones finales surgen de unir las detecciones de cada cámara considerándolos como conjuntos disjuntos.

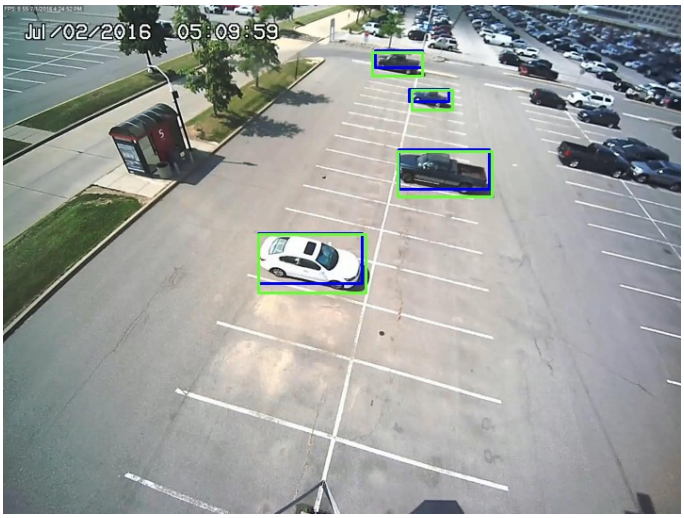


Figura 3-10. Detecciones originales y transferidas.

En la Figura 3-10 se ilustra un ejemplo de la combinación de las detecciones, representan en verde las detecciones propias, en azul las transferidas.

Ambos algoritmos, DPM y *Faster* R-CNN son usados para determinar si al añadir esta información contextual se consiguen mejores resultados en entornos multi-cámara. Los resultados obtenidos, que se presentan en el siguiente capítulo, confirman la suposición de que la combinación de información proporciona mejores resultados que los obtenidos de procesar las cámaras por separado. No se ha hecho uso del algoritmo ACF expuesto en la sección 2.3.1, ya que no cumplía con las necesidades de nuestro sistema al detectar el techo de los vehículos en vez del vehículo entero, lo que no permite proyectar el BB en el plano del suelo.

4 Experimentos y resultados

Este capítulo aborda los experimentos llevados a cabo durante este trabajo. Se hace una descripción del marco de evaluación sobre el que se basa este trabajo, detallándose el *dataset* utilizado y las métricas de evaluación empleadas. Finalmente, se presentan y analizan los resultados obtenidos y se compararán con modelos anteriores.

4.1 Marco de evaluación

Esta sección está dividida en tres partes, en la primera de ellas se describe el *dataset* empleado, en la segunda se detalla una revisión del GT realizada al comienzo de este trabajo y una tercera en la que se exponen las métricas usadas con el fin de evaluar los resultados obtenidos.

4.1.1 Dataset

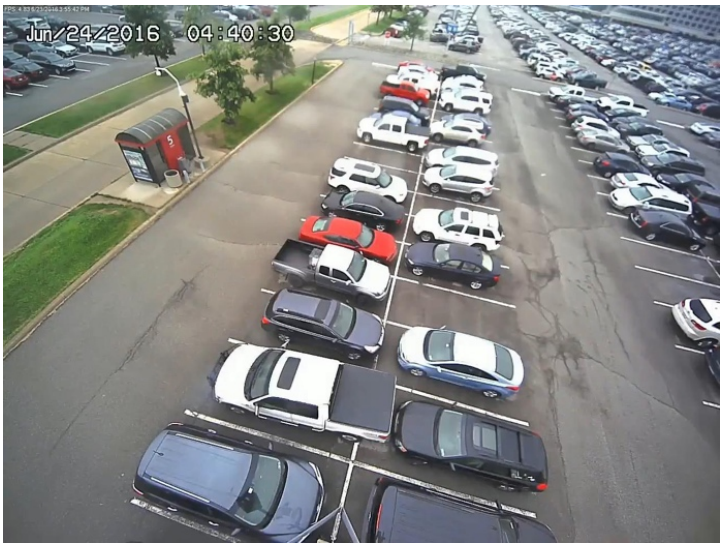
El *dataset* usado es el mismo que usa [2]. Se trata de una secuencia de imágenes grabadas en un entorno real, concretamente en el parking del Aeropuerto Internacional de Pittsburgh. De esta forma se proponía trabajar con un entorno lo más real posible. Cada *frame* tiene una resolución de 1280x960 píxeles y han sido grabados por una cámara *Panasonic WV-SW155*. Todo ello conforma el Parking Lot dataset (PLds). La Figura 4-1 muestra un ejemplo de cada uno de los puntos de vista de las dos cámaras.

El conjunto de imágenes usado para la evaluación comprende 100 *frames* de cada cámara sincronizados entre sí. En la Tabla 4-1 se detalla el conjunto de las imágenes usadas.

Nombre de secuencia		*#Frames*	#Vehículos
TEST	Multicamera_QRIDR	100	751
	Multicamera_VXUSD	100	749

Tabla 4-1. Propiedades del conjunto de imágenes del *dataset.*

a



b

Figura 4-1. Ejemplo visual de *frames* del *dataset.*

En (a) se muestra un fotograma de ejemplo de la secuencia desde el punto de vista de cámara 1 y (b) un fotograma de ejemplo de la secuencia desde el punto de vista de la cámara 2.

Además de las imágenes grabadas, todos los vehículos que aparecen en estas imágenes fueron anotados manualmente, generándose así el GT utilizado. El *dataset* y el GT están disponibles (http://www-vpu.eps.uam.es/DS/PLds/).

4.1.2 Revisión del Ground Truth

Al comienzo de este trabajo, con el fin de garantizar unos mejores resultados y más fiables, se propuso una revisión concienzuda del GT descrito en la sección anterior. Se incluyeron nuevas anotaciones de vehículos que resultaban estar ocultos ante la vista de una cámara, pero no de la otra. Concretamente, esto ocurría con los vehículos más alejados a las cámaras. Al haber añadido nuevos BB para vehículos ocluidos y lejanos se aprecia como los resultados son ligeramente peores debido a que este nuevo GT es más completo y riguroso, habiendo sido generado haciendo uso de la información de ambas cámaras simultáneamente.

Las curvas PR de la evaluación antes de la revisión del GT y después de ésta se muestra en la Figura 4-2 para la cámara 1 y en la Figura 4-3 para la cámara 2, El AUC para cada una de estas curvas se detalla en la Tabla 4-2.

Algoritmo	Cámara 1		Cámara 2	
	DPM	Faster R-CNN	DPM	Faster R-CNN
Antes de revisión	0,610	0,726	0,670	0,871
Tras revisión	0,585	0,686	0,642	0,830
Ganancia (Δ%)	4,1	5,5	4,1	4,7

Tabla 4-2. AUC antes y después de revisar el GT.

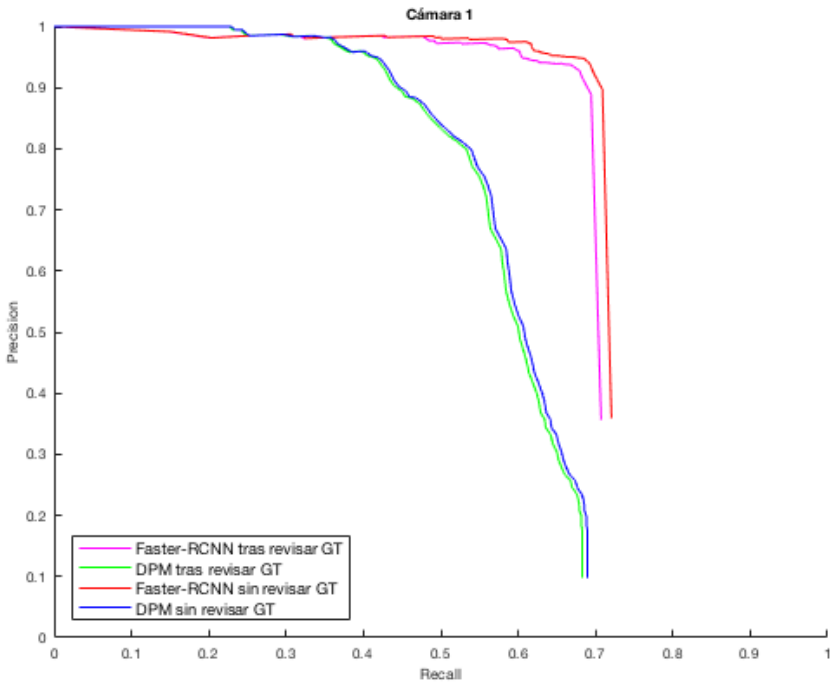


Figura 4-2. Curvas PR, GT del modelo anterior vs GT revisado en Cámara 1.

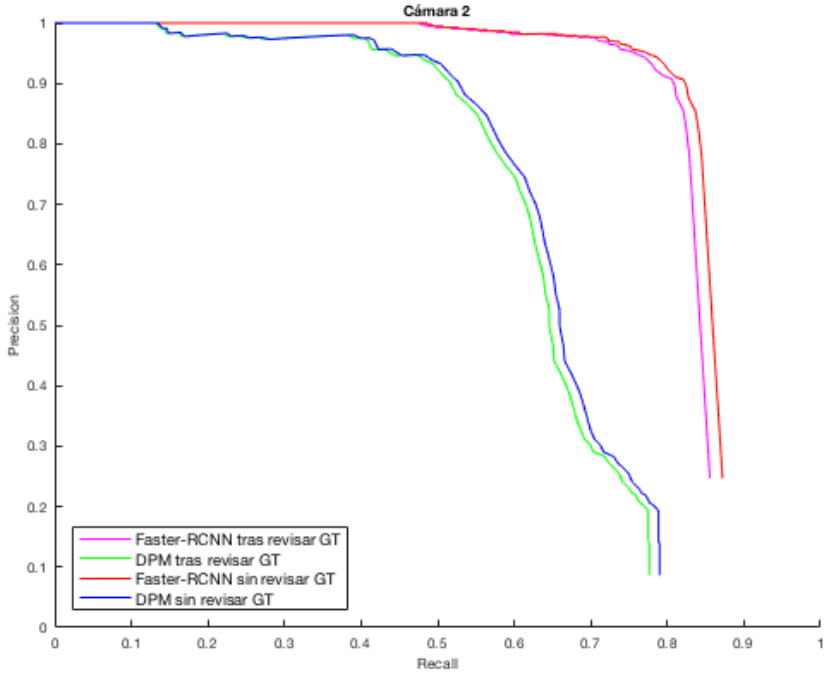


Figura 4-3. Curvas PR, GT del modelo anterior vs GT revisado en Cámara 2.

4.1.3 Métricas de evaluación

Para poder evaluar el trabajo realizado se cuantifican los resultados obtenidos. El rendimiento normalmente se mide respecto de los términos de curvas de *Precision-Recall* (PR). En estas curvas se compara las similitudes entre lo obtenido a la salida y los BB del GT. Además, para evaluar también la localización precisa del objeto detectado se tienen en cuenta tres criterios de evaluación definidos en [22], según esto, se pueden comparar hipótesis a distinta escala: distancia relativa (dr, distancia entre los centros de los BB), *cover* (toda la parte de un BB que queda cubierta por la de otro BB) y *overlap* (zona de un BB que no queda cubierta por otro BB), ver Figura 4-4. Se considera que una detección es positiva cuando la dr ≤ 0.5 o el *cover* y el *overlap* se encuentran por encima del 50 %.

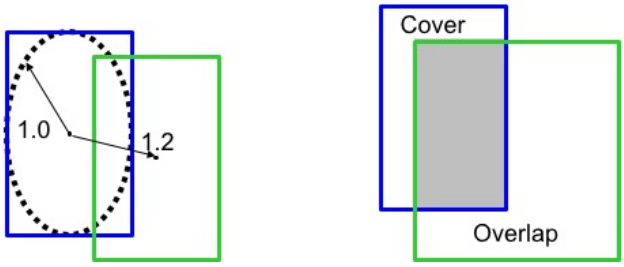


Figura 4-4. Ejemplo de métricas utilizadas.

Las curvas mencionadas antes se obtienen según estas fórmulas:

$$Precision = \frac{\#Verdaderos\ positivos}{\#Verdaderos\ positivos + \#Falsos\ positivos}$$

$$Recall = \frac{\#Verdaderos\ positivos}{\#Verdaderos\ positivos + \#Falsos\ negativos}$$

Cada uno de los términos significa:

- Verdaderos positivos: la detección es verdaderamente un vehículo, esto es, hay una coincidencia con alguna detección del GT.
- Falsos positivos: la detección no es un vehículo, esto quiere decir, no hay una coincidencia con ninguna detección del GT.
- Falsos negativos: estas son las detecciones fallidas, es decir, la detección nos indica que no es un vehículo, sin embargo, la detección del GT sí lo indica.

Además, se hace uso del área bajo la curva, *Area Under Curve*, AUC; con ella estimaremos mejor la eficiencia de los algoritmos de detección utilizados.

4.2 Resultados y comparativa con modelos anteriores

Como parte significativa del trabajo realizado se encuentran los resultados obtenidos de la revisión del GT expuestos en la sección 4.1.2. Esto se lleva a cabo para cuando se ha utilizado la máscara (ver Figura 3-3) mencionada anteriormente. Los resultados de las detecciones son evaluados con los algoritmos descritos en la sección 2.3, menos con el algoritmo ACF, el cual fue descartado por no satisfacer las características de detección necesarias, según lo expuesto al final de la sección 3.3.3.

A continuación, evaluamos los resultados arrojados por el detector de vehículos tras haber añadido la información contextual. Como ya se ha explicado en la sección 3.3.3, los criterios de umbralización utilizados para la combinación de los BB es del 50 % y del 100 % del *overlap* y del *cover.* En primer lugar, determinamos que los resultados obtenidos tras añadir la información contextual son mejores que los resultados del modelo anterior para cada uno de los algoritmos utilizados: DPM y Faster R-CNN. Lo que se muestra en la Figura 4-5 es la comparativa entre las detecciones del sistema anterior con las detecciones con la fusión de información de la cámara 2 a la cámara 1.

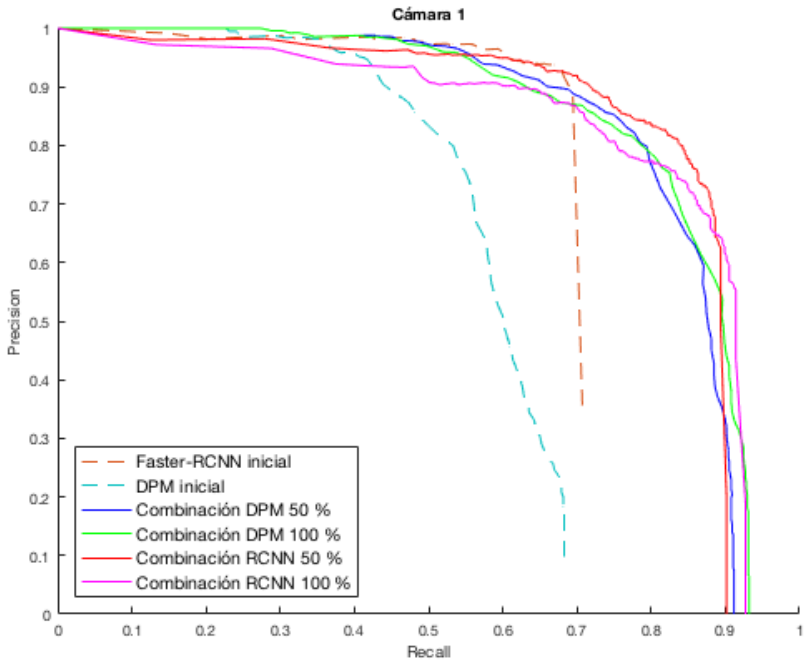


Figura 4-5. Cámara 1 con información fusionada de la cámara 2.

Lo que en la Figura 4-6 se muestra es la comparativa entre las detecciones del modelo anterior frente a las detecciones de la cámara 2 tras haber fusionado la información que proporciona la cámara 1. Se aprecia una mejora muy significativa respecto al modelo anterior teniendo en cuenta los dos umbrales utilizados para la combinación de los BB.

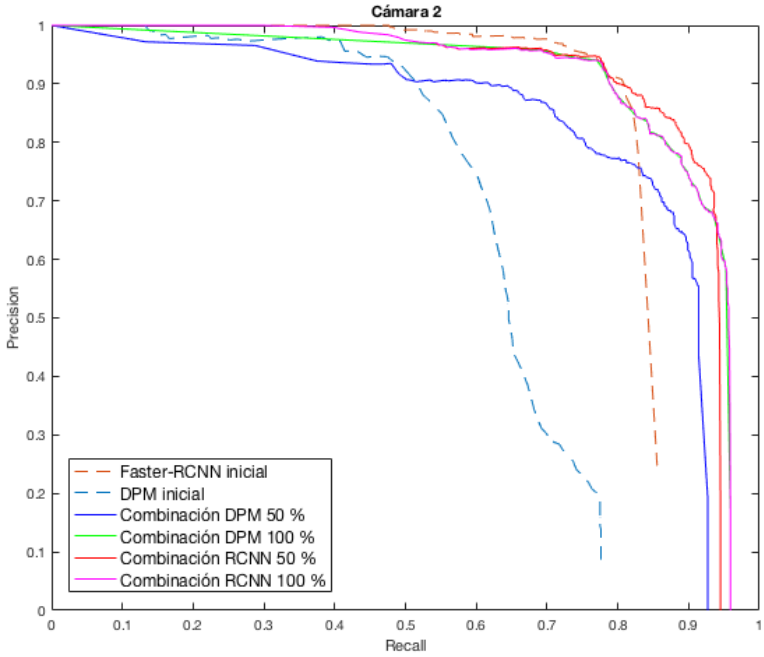


Figura 4-6. Cámara 2 con la información fusionada de la cámara 1.

En la
Tabla 4-3 se detalla el AUC obtenida para cada una de estas curvas de PR.

Algoritmo	*Cover* y *Overlap*	Cámara 1	Cámara 2
DPM	Modelo anterior	0,610	0,670
	50 %	0.831	0.841
	Ganancia (Δ%)	36,2	25,5
	100 %	**0.840**	0.899
	Ganancia (Δ%)	37,7	34,2
Faster R-CNN	Modelo anterior	0,726	0,871
	50 %	0.839	0.903
	Ganancia (Δ%)	15,6	3,7
	100 %	0.824	**0.905**
	Ganancia (Δ%)	13,5	3,9

Tabla 4-3. AUC del modelo anterior y tras fusión de información contextual.

La ganancia se obtiene restando el valor original al valor obtenido y dividiendo el resultado entre el valor original, esto nos permite conocer el porcentaje de mejora con respecto al valor original.

4.3 Discusión

De los resultados anteriores se puede afirmar que, tras la fusión de la información contextual para cada una de las cámaras respectivamente y tras la evaluación con los algoritmos DPM y Faster R-CNN, se obtienen mejores resultados en cuanto al recall y al AUC. Se observa numéricamente en el incremento del AUC reflejado en la Tabla 4-3.

Respecto al detector Faster R-CNN, se aprecia una mejora significativa del *recall*, lo que demuestra la eficiencia de este algoritmo, si bien es verdad que la *precisión* ha sufrido una ligera caída, aunque se debe tener en cuenta el punto de partida en el cual este parámetro era bastante alto, a costa de esta pérdida se ha incrementado la calidad de las detecciones. Esto ocurre también en la cámara 2, en la que la precisión del algoritmo DPM con un umbral del 50 % tras haber hecho la combinación de la información contextual proveniente de la cámara 1, baja significativamente respecto de los niveles iniciales, sin embargo, se puede apreciar una mejora muy notable en cuanto al *recall.*

Con todo ello, tras haber perfeccionado el GT, haciéndolo más riguroso, gracias a la información tomada de las dos cámaras, y haber llevado a cabo la implementación de la fusión de la información contextual y su asociación con las detecciones dadas por los algoritmos de detección, es razonable decir que se ha hecho una aportación significativa para mejorar las detecciones de vehículos del sistema de detección de vehículos desarrollado en la sección 3.2.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

En este trabajo se presentan las mejoras realizadas sobre un sistema para la gestión de las plazas libres y ocupadas de un parking. Este sistema determina la disponibilidad de las plazas haciendo uso de un detector de vehículos y mapeando cuáles de ellas se encuentran libres u ocupadas. Las mejoras introducidas buscan añadir más eficiencia en el proceso de detección mediante el uso de la información contextual. Este sistema fue diseñado para que las propias cámaras de seguridad actúen como tal y, además, sirvan para el propósito antes expuesto. Por ello, mediante la combinación de la información proporcionada por dos cámaras, obtenemos una mejora significativa en cuanto a las detecciones de vehículos se refiere, siendo este el propósito principal de este trabajo. El sistema del que partíamos hacia frente a condiciones complicadas del escenario con el que trabajaba, tales como lluvia, nieve y oclusiones totales, las cuales limitaban y reducían la eficiencia de éste. Estas mismas condiciones se han mantenido y trabajado con ellas durante este trabajo.

5.2 Trabajo futuro

Existen múltiples líneas de trabajo futuro para mejorar este sistema. Comenzando por el uso de nuevos detectores tales como el YOLO9000 descrito en [24]. Además, con el fin de comprobar la robustez del sistema, a este se le pueden añadir mayor número de cámaras para el procesamiento simultáneo de información contextual. Proponemos además en relación con lo anterior, combinar distintos detectores y cámara de forma simultánea.

Respecto de la combinación de la información, otra línea de trabajo futuro es la mejora en la combinación de información. En caso de usar más de dos cámaras, se propone desarrollar la asociación en función de las distancias entre esas cámaras y la confianza de las detecciones de cada punto de vista.

Por último, considerar otros modelos de objetos para detectar por el sistema. Analizar el estado del arte para determinar qué objetos están causando mayor interés para su detección en entornos multi-cámara.

Referencias

[1] Á. García-Martín and J. M. Martínez, "People detection in surveillance: classification and evaluation," IET Computer Vision, vol. 9, no. 5, pp. 779–788, 2015.

[2] R. Martín-Nieto, Á. García-Martín, A. G. Hauptmann, and J. M. Martínez. "Automatic vacant parking places management system using multicamera vehicle detection." *IEEE Trans. on Intelligent Transportat. Systems (In press)*, pages 1–12, 2017.

[3] R. J. L. Sastre, P. G. Jimenez, F. J. Acevedo, and S. M. Bascon, "Computer algebra algorithms applied to computer vision in a parking management system," *International Symposium on Industrial Electronics*, pp. 1675–1680, 2007.

[4] N. True, "Vacant parking space detection in static images," *Projects in Vision & Learning*, pp. 1–5, 2007.

[5] Q. Wu, C. Huang, S. y. Wang, W. c. Chiu, and T. Chen, "Robust parking space detection considering inter-space correlation," *International Conference on Multimedia and Expo*, pp. 659–662, 2007.

[6] C.-C. Huang, S.-J. Wang, Y.-J. Change, and T. Chen, "A bayesian hierarchical detection framework for parking space detection," *International Conference on Multimedia and Expo*, pp. 2097–2100, 2008.

[7] H. R. H. Al-Absi, J. D. D. Devaraj, P. Sebastian, and Y. V. Voon, "Vision-based automated parking system," *International Conference on Information Science, Signal Processing and their Applications*, pp. 757– 760, 2010.

[8] C. C. Huang, Y.-S. Dai, and S. J. Wang, "A surface-based vacant space detection for an intelligent parking lot," *International Conference on ITS Telecommunications*, pp. 284–288, 2012.

[9] C. C. Huang, Y. S. Tai, and S. J. Wang, "Vacant parking space detection based on plane-based bayesian hierarchical framework," *Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 9, pp. 1598– 1610, 2013.

[10] C. C. Huang, H. T. Vu, and Y. R. Chen, "A multiclass boosting approach for integrating weak classifiers in parking space detection," *International Conference on Consumer Electronics*, pp. 314–315, 2015.

[11] M. Tschentscher, C. Koch, M. König, J. Salmen, and M. Schlipsing, "Scalable real time parking lot classification: An evaluation of image features and supervised learning algorithms," *International Joint Conference on Neural Networks*, pp. 1–8, 2015.

[12] C. C. Huang and H. T. Vu, "A multi-layer discriminative framework for parking space detection," *International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2015.

[13] H. Xie, Q. Wu, B. Chen, Y. Chen, and S. Hong, "Vehicle detection in open parks using a convolutional neural network," *International Conference on Intelligent Systems Design and Engineering Applications*, pp. 927–930, 2015.

[14] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2013.

[15] R. B. Girshick, "Fast R-CNN," *International Conference on Computer Vision*, pp. 1440–1448, 2015.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[17] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pp. 1532-1545, 2014.
[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models" IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627-1645, 2010.
[19] R. Martín-Nieto, A. Miguélez-Sierra, A. García-Martín, and J. M. Martínez. "Improving multicamera people detection using contextual information." *Under Review*, 2018.
[20] Xiaogang Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3– 19, 2013.
[21] Thiago T. Santos and Carlos H. Morimoto, "Multiple camera people detection and tracking using support integration", *Pattern Recognition Letters*, vol. 32, no. 1, pp. 47–55, 2011.
[22] Kyungnam Kim and Larry S. Davis, "*Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering"*, pp. 98–109, Springer Berlin Heidelberg, 2006.
[23] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes", *In proc. of Computer Vision and Pattern Recognition*, pp. 878– 885, 2005.
[24] J. Redmon and A. Farhadi. "YOLO9000: Better, Faster, Stronger." In IEEE Conference on Computer Vision and Pattern Recognition conference, pages 6517-6525, 2017.

Glosario

ACF	Aggregate Channel Features
AUC	Area Under Curve
BB	Bounding Box
DPM	Deformable Parts Model
FC	Fully Connected
GT	Ground Truth
PLds	Parking Lot dataset
PR	Precision-Recall
RCNN	Region-based Convolutional Network
RD	Relative Distance
ROI	Region of Interest
RPN	Region Proposal Network
SVM	Support Vector Machine
YOLO	You Only Look Once (Detection Algorithm)